ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

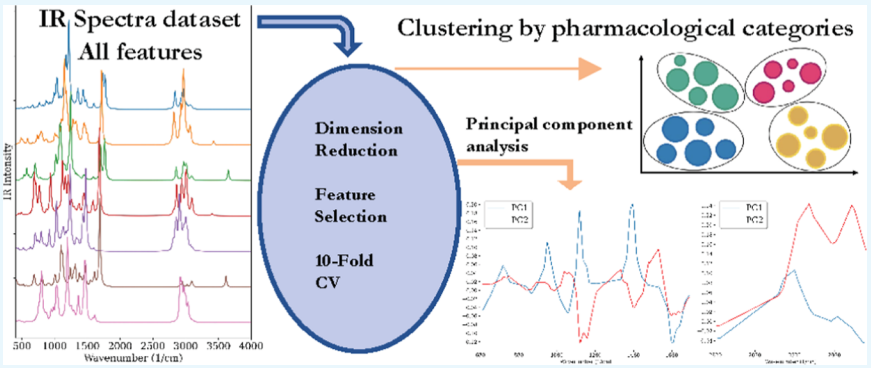
Filter Feature Selection for Unsupervised Clustering of Designer Drugs Using DFT Simulated IR Spectra Data

Kedan He*

Cite This: *ACS Omega* 2021, 6, 32151−32165

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The rapid emergence of novel psychoactive substances (NPS) poses new challenges and requirements for forensic testing/analysis techniques. This paper aims to explore the application of unsupervised clustering of NPS compounds' infrared spectra. Two statistical measures, Pearson and Spearman, were used to quantify the spectral similarity and to generate similarity matrices for hierarchical clustering. The correspondence of spectral similarity clustering trees to the commonly used structural/pharmacological categorization was evaluated and compared to the clustering generated using 2D/3D molecular fingerprints. Hybrid model feature selections were applied using different filter-based feature ranking algorithms developed for unsupervised clustering tasks. Since Spearman tends to overestimate the spectral similarity based on the overall pattern of the full spectrum, the clustering result shows the highest degree of improvement from having the nondiscriminative features removed. The loading plots of the first two principal components of the optimal feature subsets confirmed that the most important vibrational bands contributing to the clustering of NPS compounds were selected using non-negative discriminative feature selection (NDFS) algorithms.

INTRODUCTION

New psychoactive substances (NPS), also known as designer drugs, are compounds that alter the molecular structure of existing controlled substances to mimic their pharmacological effects and circumvent legislation.[1,2] According to the United Nations Office on Drugs and Crime (UNODC), as of December 2020, 126 countries had reported a total of more than 1,047 NPS.[1] Forensic analysis of NPS faces challenges such as diverse samples of unknown nature, an insufficient quantity of evidence, the need for protecting the integrity of materials for criminal investigations and legal disputes, and the demand for in-field testing. Nondestructive, low-cost, and relatively easy-to-use vibrational spectroscopy techniques such as infrared (IR) and Raman are used to characterize the structure of organic molecules.[3−6] The most common method of spectral identification is library search, in which an unknown sample is compared to each spectrum in the library, and a list of the best hits is returned based on a similarity metric.[7] The quality of library reference spectra and the robustness of similarity metrics limit the quality of library searches. The library must be large enough to contain spectra of samples similar to the unknown compound, and there must be a high degree of structural similarity between the "unknown" and the library substances in order to identify the "unknown" compound with confidence. The rapidity of emergence and the often-transient nature of some NPS compounds make it difficult to obtain a comprehensive spectral library. An alternative approach to identify never-before-seen NPS is to classify them based on structural similarity, as structurally similar compounds are likely to exhibit similar biological activities and spectroscopic characteristics.[8−10] However, the increasing complexity and diversity of NPS compounds prevent systematic classification with respect to their structural similarity by visual inspection alone. The similarity of chemical structures could be quantified

Received: September 7, 2021
Accepted: November 1, 2021
Published: November 16, 2021

using their 2D/3D molecular fingerprint representation to calculate the Tanimoto coefficient. Conversely, representing spectra as a linear vector of intensities allows quantitative comparison using statistical correlation coefficients.

Pattern recognition leverages information extracted from training samples to assign an unknown sample to a given class or category. Hierarchical clustering analysis is an unsupervised technique that provides multilevel nested results that can be used to help guide the identification of drug compounds that share a common structural and spectral feature. The IR spectrum is represented as a vector in a multidimensional space, where each dimension (feature) corresponds to a certain wavenumber and the corresponding absorbance (intensity). However, with the existence of a large number of features, learning models tend to overfit, and their learning performance degenerates. It has been verified that for complex analytical systems like vibrational spectroscopy data, it is very important and essential to conduct feature selection to gain better prediction performance.[11] Dimension reduction approaches use a low-dimensional space to substitute the original high-dimensional variable space. For example, projection methods, such as principal component analysis (PCA) and partial least squares (PLS), are used to reduce the impact of collinearity, band overlaps, and redundant noise irrelevant to the property of interest by replacing the original variables with a few latent variables or principal components (PCs) of larger variance.[12,13] However, the latent variables are hardly interpretable compared to original variables. In contrast, feature selection is based on the assumptions of choosing a small number of variables that can improve the prediction performance and provide easier interpretation. Unsupervised feature selection methods do not utilize label information and can be classified into the filter model, wrapper model, and hybrid model according to different selection strategies. Filter feature selection algorithms are computationally efficient, as they evaluate the relevance of a feature using certain statistical criteria and are independent of any clustering algorithm. A wrapper model evaluates the candidate feature subsets by the quality of clustering and is more biased to the chosen clustering algorithm. To alleviate the computational costs and benefit from the efficient filtering criteria, the hybrid model bridges the gap between the filter and wrapper models by utilizing filtering criteria to select the candidate feature subsets and then evaluates the quality of clustering of each candidate subset.[14] The subset with the highest clustering quality will be selected.

The underlying idea of ensemble feature selection is combining the subsets of several individual feature selection methods (feature selectors) to obtain better or comparable results rather than using a single feature selection approach. When the data dimensionality is very high but the number of samples is relatively small, ensemble feature selection is used to improve the stability. A more appropriate (stable) feature subset is obtained by combining the multiple feature subsets of the ensemble, as the aggregated result tends to obtain more accurate and stable results, reducing the risk of choosing an unstable subset. The other main motivation is to increase the diversity: different feature selectors provide different enough outputs on the same sample of data and decrease the chance of inaccurate prediction of samples. The main issues involved in the process are (1) the individual feature selection methods to be used; (2) the number of different feature selection methods to use; and (3) the aggregation method for feature subset generation. Ensemble feature selection can typically be categorized into the combination of labeled predictions, the combination of subsets of features, and the combination of ranking of features, which depends on whether the feature selector returns a subset of relevant features or an ordered ranking of all the features according to their relevance. When filter methods are used which return an ordered list of all features, a threshold must be chosen to reduce the dimensionality of the problem, which can become computationally expensive. The combined feature subsets, on the other hand, are generated by computing the intersection or the union of the ranked features. The intersection consists in selecting only those features which are selected by all the feature selectors, whereas the union consists in combining all the features which have been selected by at least one of the feature selectors. The potential issue is that it can lead to very restrictive sets of features (an empty set) or to select even the whole set of features, respectively. To alleviate the problem, a simple approach is to include a subset of ranked features into the final ensemble only if it contributes to improving the learning tasks. Lastly, the relevancy of the final selection of features needs to be evaluated, which is possible using synthetic data where label information is known.

In this study, we performed unsupervised clustering analysis of a set of the most common NPS compounds whose IR spectra were simulated using density functional theory (DFT). We compared the correspondence of hierarchical clustering of NPS compounds into structurally distinct groups using 2D and 3D binary molecular fingerprints with cluster labels assigned according to generally accepted chemical/pharmacological classifications. Similarly, the spectral similarity can be quantified by statistical measurements such as Pearson and Spearman correlation coefficients. The clustering performance was quantified using the silhouette score,[15] adjusted rand index (ARI),[16] and normalized mutual information (NMI).[17] Four filter-based feature selection methods developed for clustering tasks were explored in this study: spectral feature selection (SPEC),[18] laplacian score (LS),[19] unsupervised discriminative feature selection (UDFS),[20] and non-negative discriminative feature selection (NDFS).[21] The class distributions in the NPS dataset are highly imbalanced. The oversampling technique, synthetic minority oversampling technique (SMOTE),[22] was implemented in the feature selection process, and its efficiency in improving the clustering performance of imbalanced datasets was examined. Finally, aggregated feature subsets were generated using a fusion-based ensemble technique. The optimal feature subset of IR spectroscopy for NPS compound clustering was identified. When comparing the loading plots of the first two PCs of the full range and dimension-reduced datasets, it can be confirmed that the most discriminative features are retained even after the feature reduction.

RESULTS AND DISCUSSION

Comparison with Experimental Gas IR Spectra Using Quantitative Correlation Measurements. To assess the quality of the DFT-simulated IR spectra and the dependence on the basis sets, correlation coefficients are calculated from the comparison of DFT and experimental spectra. Usually, the scaling factors are derived by minimizing the rmsd of peak wavenumbers; however, the exact combination of DFT/6-31++G(d, p) and DFT/6-311++G(d, p) is not available in the CCCBDB[30] database; hence, the same scaling factor 0.964 for DFT/6-31+G(d, p) was used for the other two larger basis sets. The optimal scaling factors were determined by maximization of the correlation coefficient to assess the usefulness of using two

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

Table 1. Average Correlation Coefficient of DFT Spectra Compared to Experimental Data Shown in Figure 1

	Pearson r			Spearman ρ		
basis set	(unscaled)	(scaled)[a]	optimized (scaling factor[b])	(unscaled)	(scaled)[a]	optimized (scaling factor[b])
6-31G(d)	0.455	0.723	0.784 (0.967)	0.739	0.859	0.863 (0.964)
6-31+G(d, p)	0.497	0.676	0.798 (0.965)	0.676	0.860	0.865 (0.967)
6-31++G(d, p)	0.497	0.676	0.798 (0.965)	0.763	0.860	0.865 (0.967)
6-311++G(d, p)	0.549	0.643	0.798 (0.977)	0.779	0.857	**0.866** (0.971)

[a]Scaling factors 0.960 and 0.964 are used for 6-31G(d) and 6-31+G(d, p), respectively, as reported in the CCCBDB database.[24] For larger basis sets, the scaling factor 0.964 was applied. [b]The optimized scaling factors were obtained from maximization of the statistical correlation coefficients. The scaling factor 0.971 was used for generation of the IR spectral dataset using the 6-311++G(d, p) basis set.

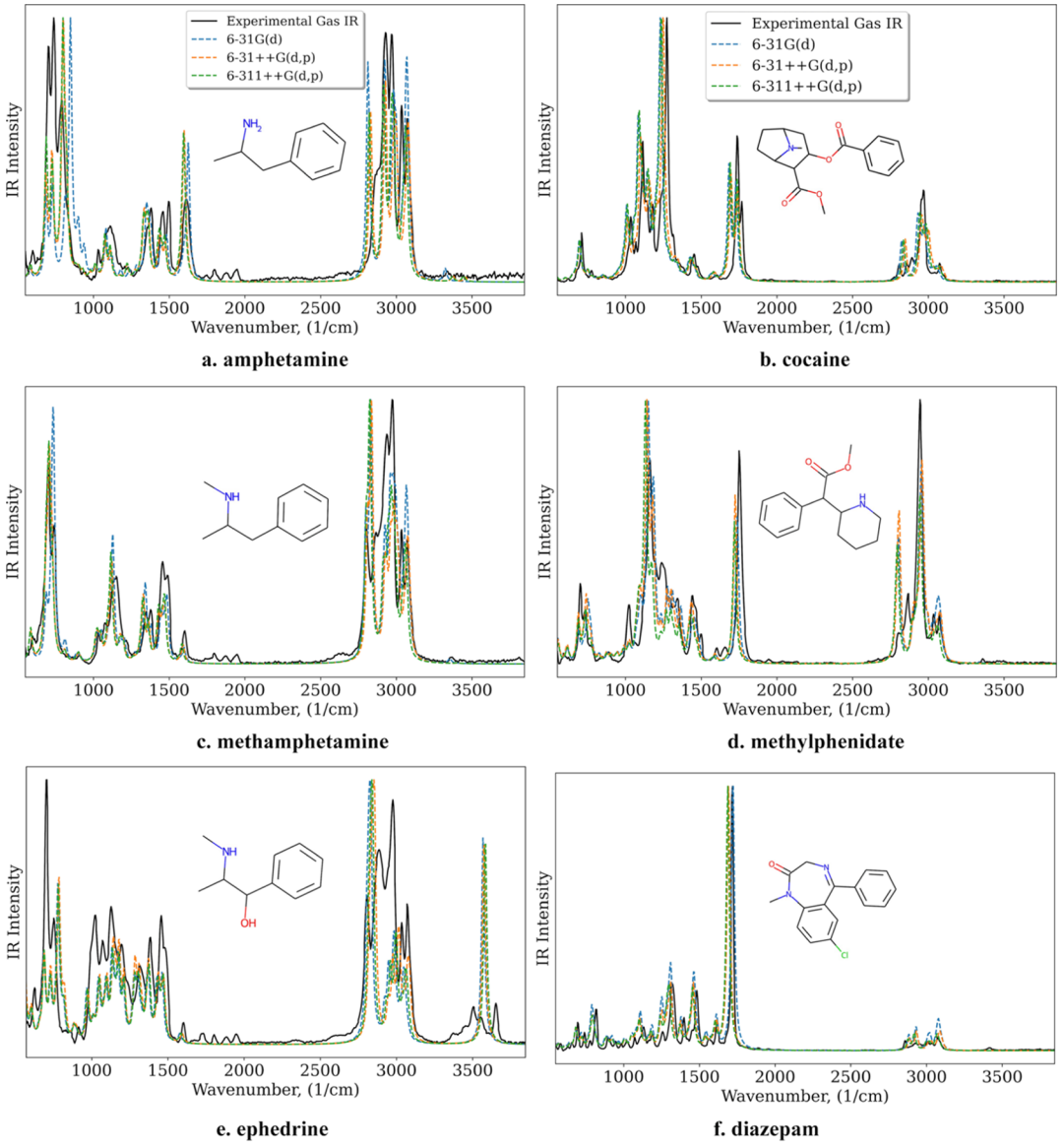


Figure 1. IR spectra of the lowest-energy conformers calculated by DFT using different basis sets in comparison with the experimental spectra. (a) Amphetamine, (b) cocaine, (c) methamphetamine, (d) methylphenidate, (e) ephedrine, and (f) diazepam.

correlation coefficients as a quantitative measure of spectral similarity. Table 1 summarizes the average results for six compounds, for which the experimental gas IR spectra are available at the NIST.[23] Figure 1 shows a visual spectral comparison of the IR spectra of the lowest-energy conformer of each compound. Spearman's correlation coefficient consistently gives a comparatively higher score with or without using the scaling factors. In terms of reproduction of the experimental spectra, the unscaled 6-311++G(d, p) spectra resulted in the highest spectral correlation coefficients (r = 0.549 and ρ = 0.779). Applying scaling factors resulted in a somewhat larger increase in Pearson, indicating that Pearson is more sensitive to the exact wavenumbers of the peak position. The optimization of both statistical measures gave rise to scaling factors that only differ insignificantly from the literature and follow the general trend that the larger the basis set, the higher the scaling factor

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

Table 2. Correlation Coefficients of Spectra Shown in Figure 1 Using the 6-311++G(d, p) Basis Set

	Pearson r			Spearman ρ		
compounds	(unscaled)	(scaled)[a]	optimized (scaling factor[b])	(unscaled)	(scaled)[a]	optimized (scaling factor[b])
amphetamine	0.570	0.794	0.795 (0.965)	0.772	0.836	0.845 (0.975)
methamphetamine	0.657	**0.835**	0.843 (0.962)	0.803	0.858	0.866 (0.972)
ephedrine	0.473	0.575	0.612 (0.981)	<u>0.673</u>	0.866	0.873 (0.957)
cocaine	**0.796**	0.638	0.885 (0.990)	**0.859**	**0.901**	**0.909** (0.973)
methylphenidate	0.541	0.678	0.704 (0.983)	0.821	0.869	0.882 (0.974)
diazepam	<u>0.255</u>	<u>0.337</u>	**0.951** (0.981)	0.745	<u>0.813</u>	0.822 (0.973)

[a]The scaling factor 0.964 was used for DFT/6-311++G(d, p)-calculated IR spectra. [b]The optimized scaling factors were obtained from maximization of the statistical correlation coefficients for quantitative quantification of spectral similarity. The highest and lowest correlation coefficients are marked in bold and underlined, respectively.

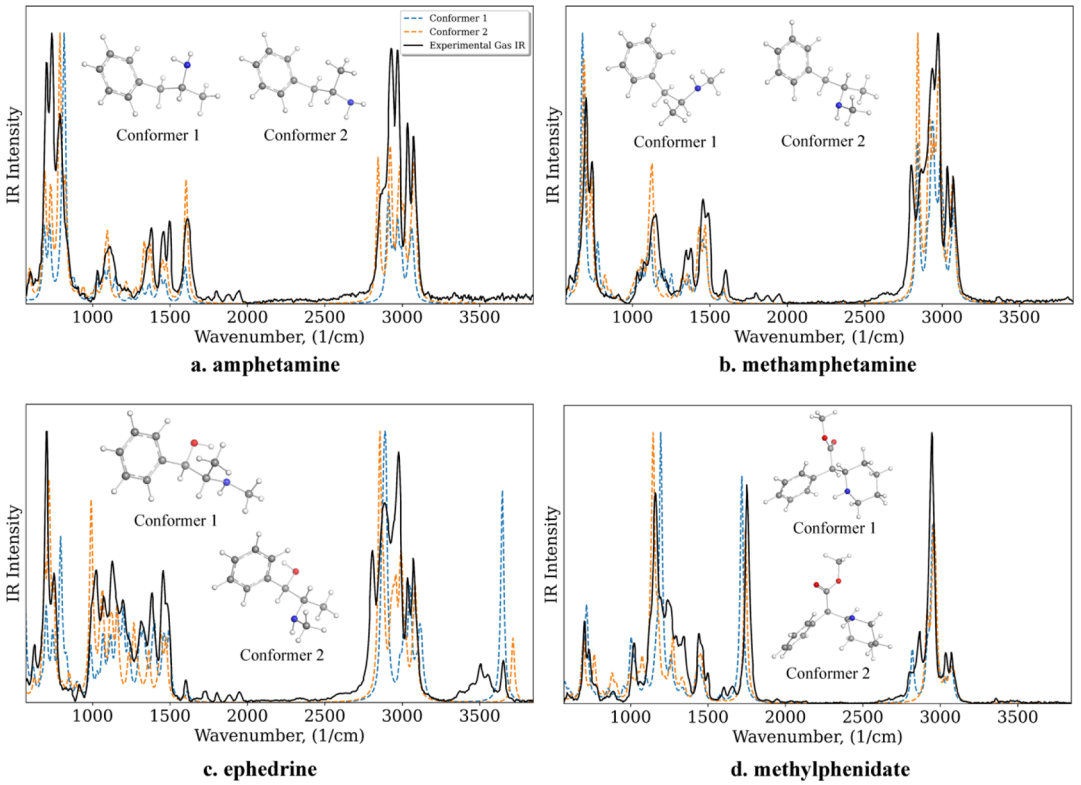


Figure 2. DFT/6-311++G(d, p) IR spectra of two conformers in comparison with the experimental spectra. (a) Amphetamine, (b) methamphetamine, (c) ephedrine, and (d) methylphenidate.

using Spearman. At the same time, application of the optimized scaling factor improves the Pearson more significantly, in contrast to the marginal improvement of Spearman. Finally, the scaling factor 0.971 was used for generation of the IR spectral dataset using the DFT/6-311++G(d, p) level of theory.

The corresponding compound-wise correlation coefficients are listed in Table 2. Diazepam stands out as an interesting case that it resulted in the largest Pearson correlation coefficient and the lowest Spearman correlation coefficient among all compounds. From Figure 1f, it can be seen that the experimental spectrum of diazepam is dominated by the carboxyl bond stretching band at around 1760 cm^{-1} followed by the benzene ring modes at around 1500 and 1350 cm^{-1}. The scaled DFT spectrum placed these bands at lower wavenumbers, and the optimization significantly increased the Pearson correlation by matching the most intense bands. It seems that reproduction of a dominant feature in the reference spectrum has a decisive impact on the Pearson correlation coefficient. Similarly, in the case of cocaine in Figure 1b, the Spearman correlation is way above the average value of all compounds for the scaled spectrum, but the Pearson correlation is lowered compared to that of the unscaled spectrum. The DFT spectrum overestimated the intensity for the C−H bending bands below 1300 cm^{-1} and underestimated the intensity for the C−H stretching bands above 2800 cm^{-1}. Overall, it suggested that the Spearman correlation coefficient could provide a better estimate of the overall similarity of the spectra, whereas the Pearson correlation coefficient is more sensitive to the peak position and intensity of the dominant features.

These two correlation coefficients also differ in terms of their sensitivity to conformational changes, as seen in Figure 2 and Table 3. Two conformers for which Pearson correlation coefficients are distinctively different are shown for four compounds, along with the optimized structures. Conformational changes usually result in changes in band intensity, as seen in Figure 2a,b. For methamphetamine, the Pearson correlation increased from 0.767 to 0.860, but the Spearman correlation stayed the same. In the case of ephedrine in Figure 2c, the

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

Table 3. Correlation Coefficients[a] of the Spectra Shown in Figure 2 Using the 6-311++G(d, p) Basis Set

	conformer 1		conformer 2	
compounds	r	ρ	r	ρ
amphetamine	0.674	0.829	0.820	0.848
methamphetamine	0.767	0.853	0.860	0.853
ephedrine	0.612	0.873	0.802	0.849
methylphenidate	0.638	0.888	0.808	0.882

[a]optimized scaling factor of each calculated IR spectrum was used.

rotation of the hydroxyl group toward the benzene ring in conformer 2 decreased the O−H stretching band intensity and right-shifted to a higher wavenumber, whereas the rotation of the amine side chain right-shifted the benzene ring modes and left-shifted the N−H bending band. As for methylphenidate in Figure 2d, the lower Pearson correlation of conformer 1 is caused by the poor match of stretching bands of the ester functional group. However, these spectral differences are not reflected when comparing the Spearman correlation coefficients of the two conformers.

Correspondence of Structural and Spectral Similarity. Clustering is used to find groups of objects that are more similar to each other than to other clusters. There are many clustering algorithms available; choosing a clustering technique is often a trial-and-error process that is very dependent on the dataset.[25−30] However, determining the best clustering algorithm for NPS IR data is outside the scope of this study, and hierarchical clustering is used for its ease of visualization of cluster relationships on a dendrogram and heatmap. The ward linkage method forms clusters by combining two clusters that result in the least increase in variance from an iterative ANOVA test.

The correspondence of the structural and spectral similarities of NPS compounds with respect to the manually assigned class label is investigated. Because conformational changes have no effect on the compound's 2D SMILES, only the lowest-energy conformer's SDF file is used to generate the 3D description in this analysis. For the spectral similarity analysis in this section, the Boltzmann distribution-weighted total IR dataset is generated using all optimized conformers of each NPS compound (see the Supporting Information for more details). The chemical structural diversity of the 127 unique NPS compounds can be characterized by the calculation of the Tanimoto similarity score for the 8001 pairs ($n(n-1)/2$). Figure 3 shows the right-skewed distributions of all structural similarities, and in contrary, the spectra similarities exhibit left-skewed distributions, with 50% of the compound pairs having a Spearman score equal to or greater than 0.759. There are also significant discrepancies among structural fingerprint approaches, with Morgan and E3FP being even more right-skewed, with mean similarity scores of 0.170 and 0.132, respectively. Figure 4 compares the use of structural and spectral similarity methods in retrieving a query compound's nearest neighbors (hits 1−5). The accuracy ratio is calculated by dividing the total number of retrieved compound pairs with the same assigned class label by the total number of compound pairs. Overall, 2D molecular fingerprints provide consistent better performance even when more top hits are chosen, whereas spectral similarity search delivers inferior and rapidly declining performance in its ability to retrieve compounds from the same class. The 3D description E3FP does not enhance the

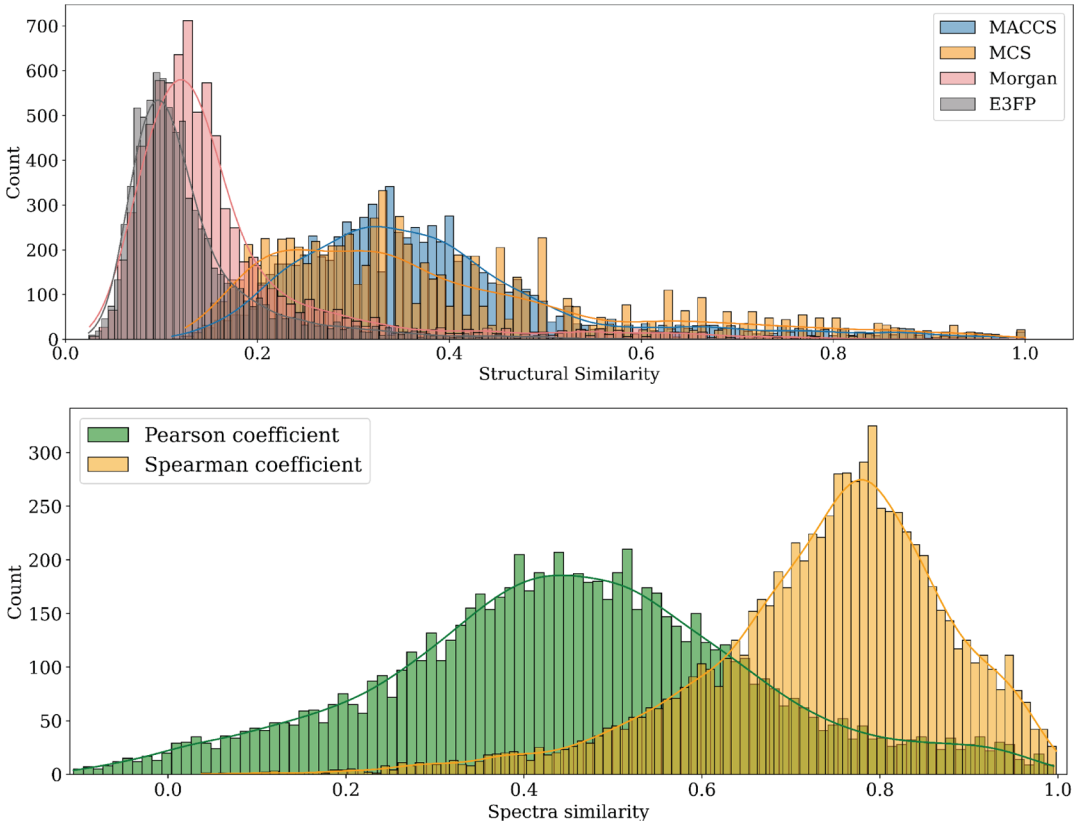


Figure 3. Frequency distribution of the Tanimoto coefficient and spectral correlation coefficients of pair-wise comparison of drug compounds.

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

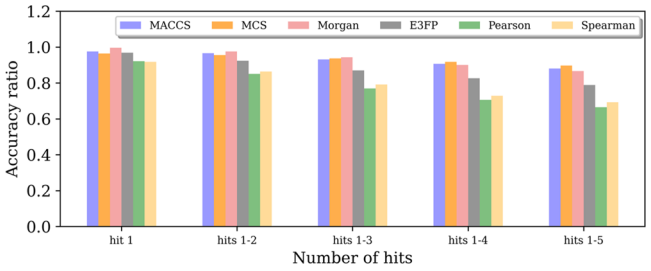


Figure 4. Accuracy ratio of retrieving compound(s) of the same class.

identification of structurally similar substances specified by commonly accepted NPS categorization.

As two examples, Figures 5 and 6 display the top hits from the spectral similarity and structural similarity searches for the query molecule THC (delta-9 tetrahydrocannabinol) and MDA (3,4-methylenedioxyamphetamine). THC is the primary component of the marijuana plant that produces psychoactive effects and a schedule II substance. The structural and spectral similarity searches return the same top-three hits: DMHP and synhexyl, synthetic analogues of THC, and CBN (cannabinol), a derivative of THC. Despite a relatively high MACCS similarity score (0.690), CBD (cannabidiol), a major nonpsychotropic constituent of cannabis, was assigned low similarity scores based on MCS (0.353) and both spectral correlation coefficients (r = 0.505 and ρ = 0.531). It is evident that when a structural change causes a significant departure of the most intense bands, the spectral correlation coefficients can distinguish the spectral differences properly. On the other hand, Spearman severely overestimates the spectral similarity and subsequently retrieves incorrect hits, as shown in Figure 6. In Figure 6a, MBDB (*N*-methyl-1,3-benzodioxolylbutanamine) and MDMA (3,4-methylenedioxymethamphetamine), two stimulants of the amphetamine family, were returned as the top-two hits using Pearson when MDA is used as the query molecule; however, these two compounds were returned as the 24th and 32nd hits using Spearman, respectively. In contrary, as seen in Figure 6b, 2C-T-2 and 2C-B were assigned the highest Spearman similarity value (ρ = 0.964 and 0.961, respectively).

Finally, all clustering trees were compared using three measures: silhouette scores based on internal proximity of information intrinsic to the data and ARI and NMI assessed by comparing clustering partitions with an external class label. Figure 7 shows the silhouette scores of Pearson and Spearman clustering trees calculated for K = 2−K = 50 clusters. The silhouette score increases as the intracluster distance decreases and the intercluster distance increases, so that the optimal number of clusters K will correspond to the highest silhouette value. The drawback of the Spearman correlation coefficient is also manifested in the highest silhouette score when dividing all samples into two clusters, as the Spearman similarity score is frequently exaggerated for many compound pairs. According to all three clustering measures shown in Figure 8, the MCS similarity clustering tree based on shared core chemical fragments outperformed all other clustering trees, followed closely by the substructural key fingerprint MACCS clustering tree. Consistent with the preceding analysis, the 3D description E3FP could not be used to sufficiently classify NPS substances based on their pharmacological/structural classification, indicated by the lowest silhouette score of 0.21. With little higher silhouette scores than those of E3FP but a lower ARI, 0.39 and 0.41, respectively, there is no significant difference between the two spectral similarity clusterings. Because Spearman correlation scores appear to be more influenced by the general pattern of the IR spectra instead of the most intense bands, this clustering should benefit more from feature selection by removing nondiscriminative features.

Feature Selection Evaluated Using Hierarchical Clustering. Figure 9 shows the scaled feature important score plots as a function of the vibrational wavenumber to assist in understanding the differences in feature subsets generated by the four feature selectors and the effect of applying the SMOTE.

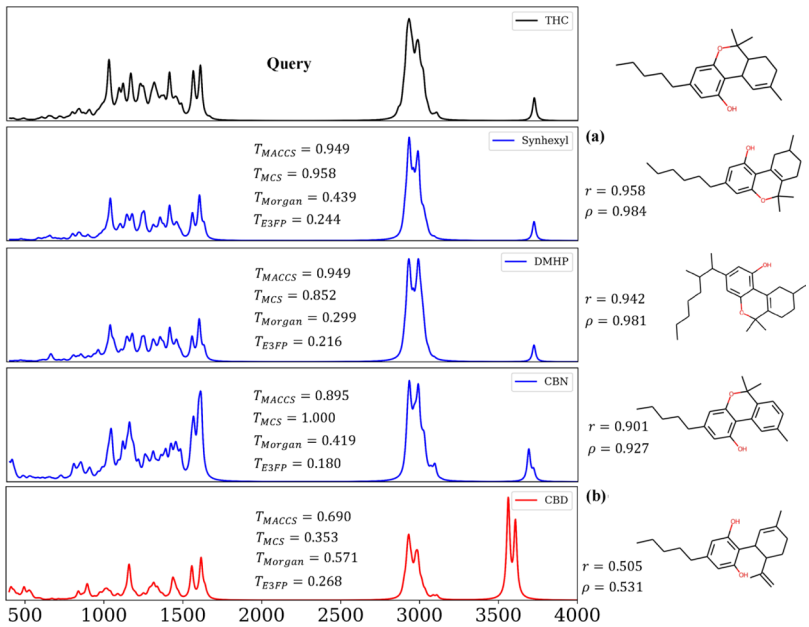


Figure 5. Search results for the query compound THC. (a) Query molecule THC and top-three hits from the spectral similarity search. (b) Spectrum of CBD and similarity scores. CBD returned as the 74th and 122nd hits from spectral similarity searches using Pearson and Spearman correlation coefficients, respectively.

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

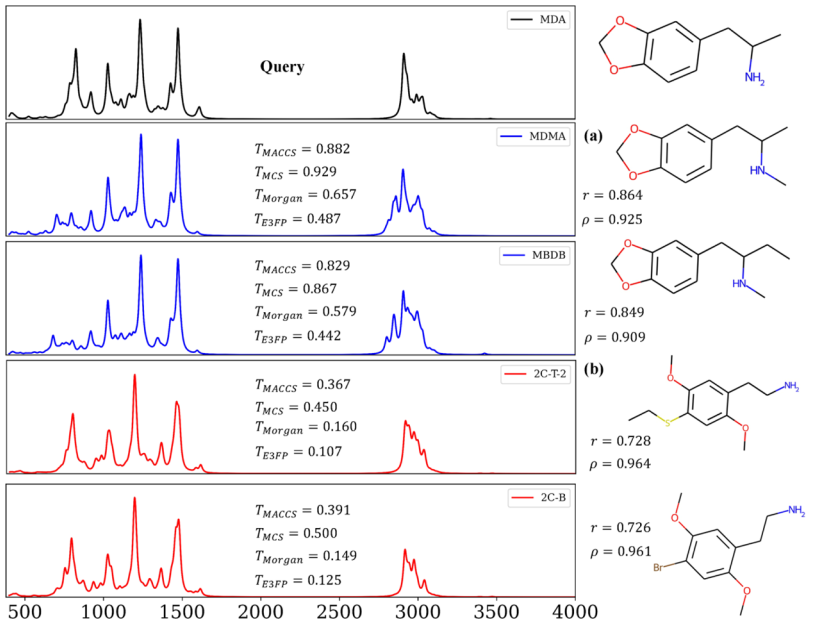


Figure 6. Search results for the query compound MDA. (a) Top-two hits using Pearson spectral similarity. (b) Top-two hits using Spearman spectral similarity.

Table 4 also lists the top-10 features chosen by all four selectors. As shown in Figure 9a,b, applying the SMOTE produces the largest differences in the feature subsets identified using SPEC and LS. For example, a feature around 3646−3654 cm^{-1} was given higher importance scores compared to when the SMOTE was not used. Another distinction is that different feature selectors analyze and sample features in varying manners. When the feature importance threshold is gradually lowered to subsequently include more top-ranking features, the first three feature selectors (SPEC, LS, and UDFS) sample a group of features in a localized fashion from one region to another. In contrary, the NDFS algorithm produces the most "sparse" selection by assigning fewer features with very high importance, resulting in a more scattered feature selection across the whole spectrum.

The clustering results of different feature selectors with similarity matrices calculated using Pearson and Spearman correlation coefficients are summarized in Figure 10. From this analysis, the following observations can be made. First, the SMOTE appears to provide no significant improvement of the clustering tasks, as the ARI changes as the number of features selected follows the same pattern for all four selectors with or without applying the SMOTE. Second, feature selection using LS and NDFS effectively reduces the feature number and improves the clustering performance in comparison to the baseline that uses the full-range features. The NDFS algorithm exploits discriminative information by evaluating features jointly resulting in a higher ARI and NMI while using the smallest feature subset. Lastly, confirming the previous analysis, Spearman clustering benefited more from using feature selectors.

The feature subsets that resulted in the highest ARI score of Spearman clustering trees using LS and NDFS algorithms are combined together by computing the union or intersection of both sets. Ensemble 2 includes the intersection of the optimal feature subsets from LS and NDFS with the SMOTE incorporated. The number of features selected in each feature subset and clustering results evaluated using the ARI and NMI are summarized in Table 5. Clustering using the ensemble 2 feature subset gives the highest ARI and NMI among all four ensembles. However, from the number of features in ensembles, it is clear that most features selected by NDFS are also selected by LS, and the ensemble feature subset clustering performance is comparable to that of the individual selectors. Given the abovementioned observation, the NDFS algorithm is able to select more informative features compared to all other filter-based feature selection models.

PCA was carried out to reduce the dimensionality of the full range (baseline, m = 1181) and the ensemble 2 subset datasets, and the loading plots of the first two PCs are shown in Figure 11. When the ensemble 2 feature subset was used, the total explained variance increased from 16.95 to 13.50% and 24.88 to 13.76% for PC1 and PC2, respectively. The same set of most important vibrational bands contributing to the clustering of NPS compounds was selected according to the PC1 and PC2 loading plots of the baseline and ensemble 2 datasets.

CONCLUSIONS

The IR spectra of NPS compounds were calculated using DFT in this study. The spectral similarity was quantified using two statistical measures: Pearson's product moment correlation and Spearman's rank correlation coefficients. When using the gas-phase experimental spectra as a reference, it is shown that Pearson is more sensitive to the intensity and peak position of the most intense bands and to the spectral changes caused by conformational changes. On the other hand, Spearman is better suited to describe the overall pattern of the full spectrum but tends to overestimate the similarity of the spectra. The ability to retrieve compounds of the same structure/pharmacological class using spectral similarity searches was evaluated and compared to structural similarity searches using 2D/3D molecular fingerprinting. Hierarchical clustering using MCS similarity proved to be a suitable method to group NPS compounds into clusters with different maximum common substructures and gave the best partition based on the ARI and NMI calculated by comparing to externally assigned class labels. The clustering

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

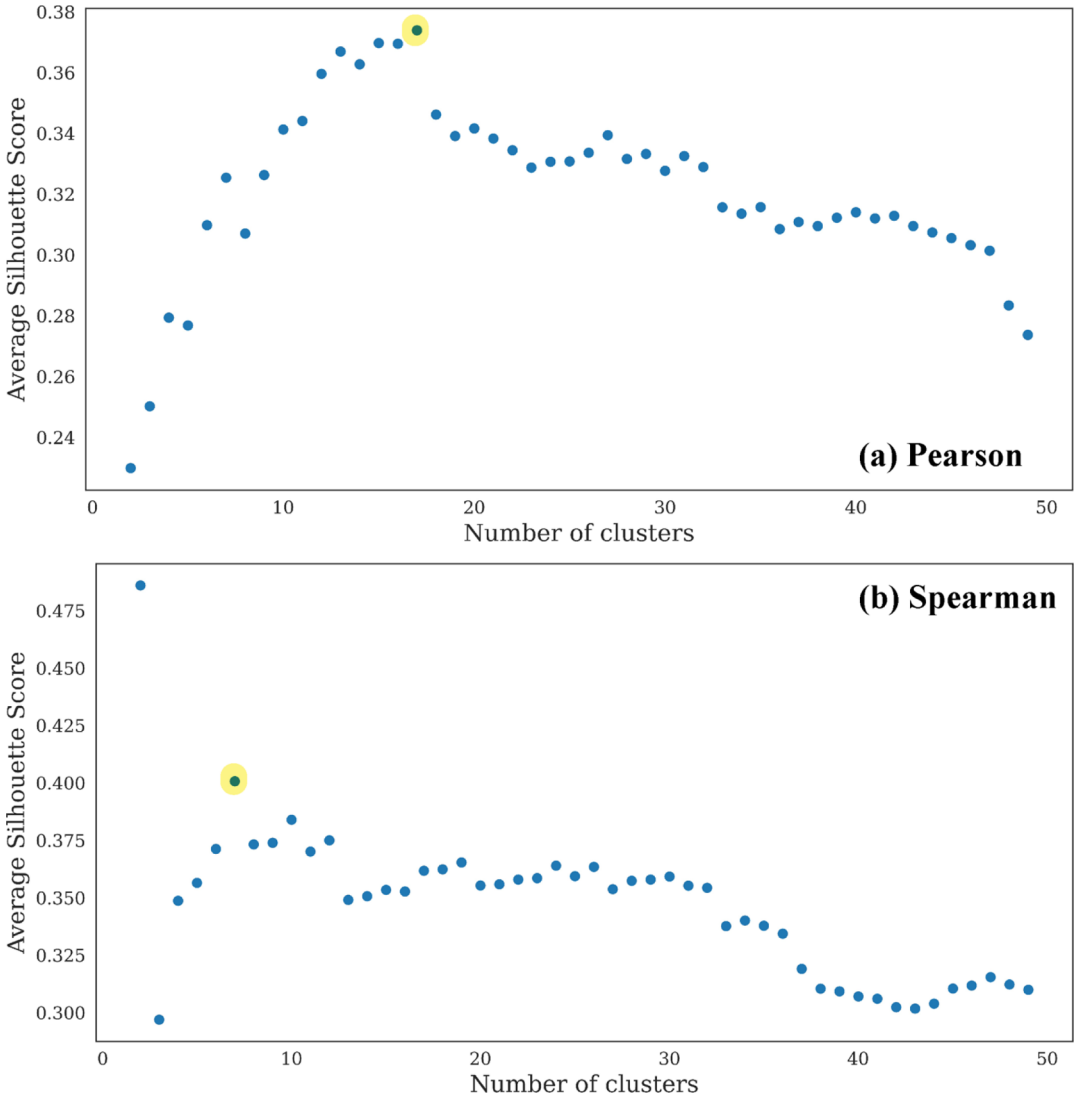


Figure 7. Silhouette scores as a function of cluster $2 < K < 50$ of spectral similarity clustering trees. (a) The optimal K is determined to be 17 using the Pearson similarity coefficient. (b) The maximal silhouette score using the Spearman similarity coefficient corresponds to $K = 2$, gradually decreasing when $K \geq 7$.

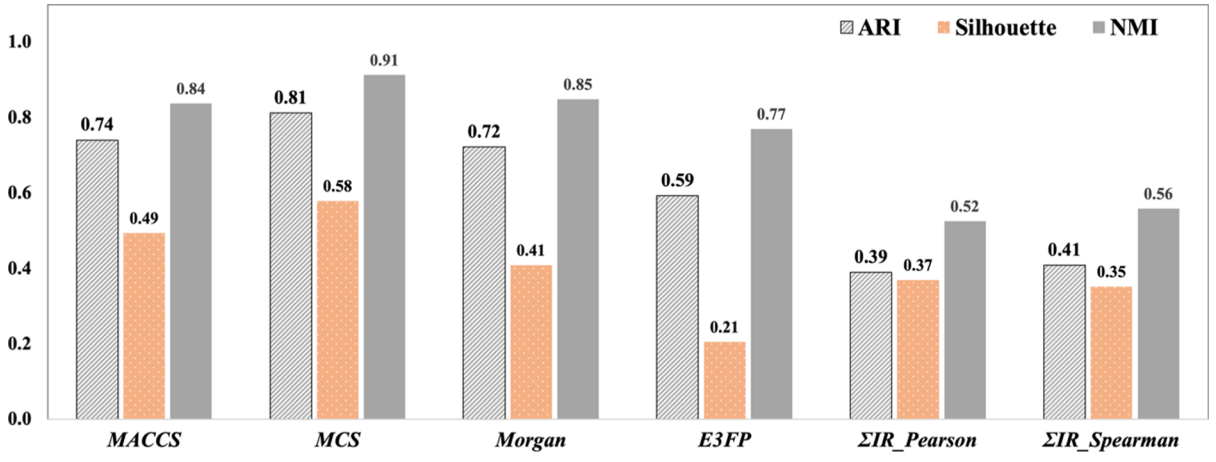


Figure 8. ARI, silhouette score, and NMI of all clustering trees evaluated using an externally assigned class label. The Boltzmann distribution-weighted total IR spectra ($\sum$IR) were used; see the Supporting Information for more details.

trees generated using the two spectral similarities showed the lowest agreement with the external class labels.

Since Spearman tends to overestimate the spectral similarity based on the overall pattern of the full spectrum, it is expected to

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

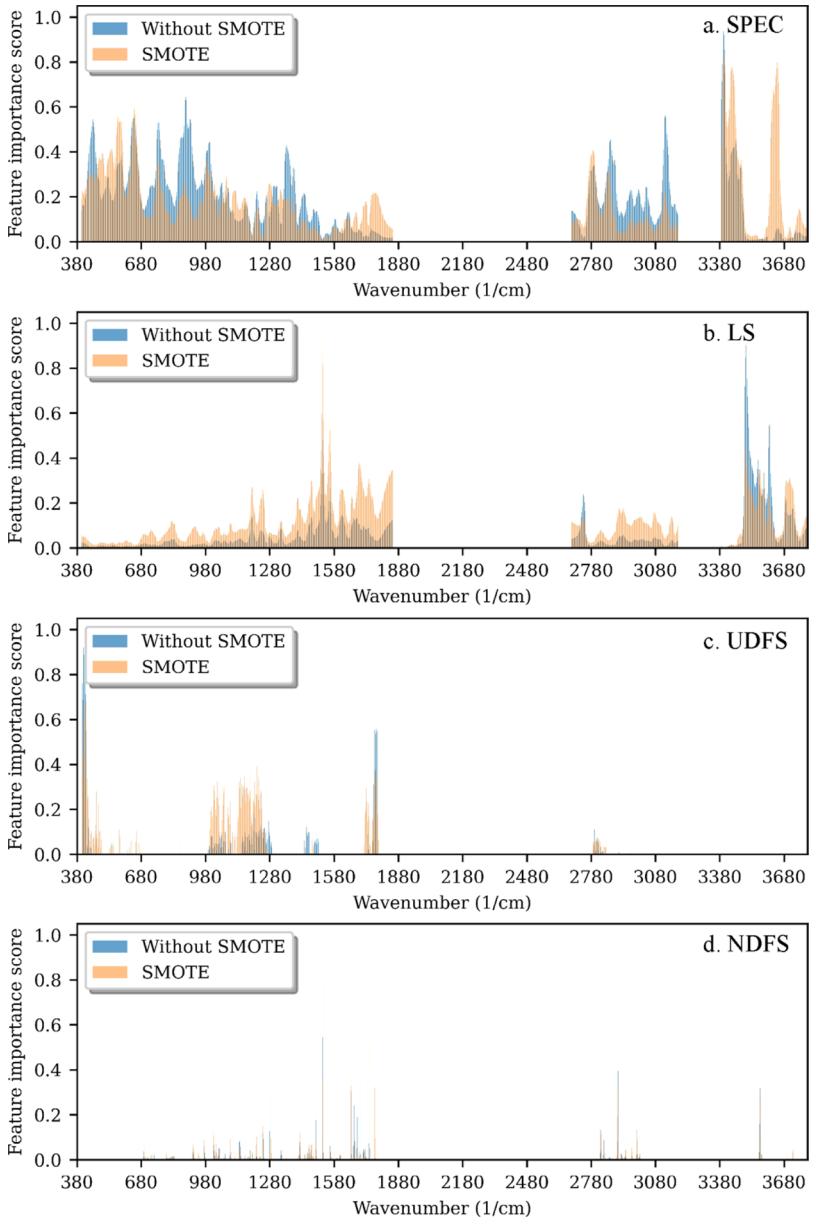


Figure 9. Scaled feature importance score from different feature selectors, with or without applying the SMOTE. Four different filter-based feature selection methods were used in calculating the feature importance score: (a) SPEC, (b) LS, (c) UDFS, and (d) NDFS.

Table 4. Top-10 Features (in cm^{-1}) Selected by Four Filter Methods

SPEC		LS		UDFS		NDFS	
w/o SMOTE	SMOTE	w/o SMOTE	SMOTE	w/o SMOTE	SMOTE	w/o SMOTE	SMOTE
3398	3400	3502	1526	410	404	1526	1526
3400	3402	3504	1524	404	422	1528	1746
3396	3648	3500	1528	412	418	1746	1658
3394	3398	3506	1522	414	416	2904	1528
3402	3650	3508	1560	402	424	1702	1660
3392	3646	3498	1530	424	408	1658	2904
3404	3404	3510	1558	418	420	3568	1770
3390	3436	3512	1562	406	414	1282	3570
888	3438	1526	3502	420	402	1660	3568
3388	3440	3612	1556	408	406	1674	2902

benefit more from feature selection to remove nondiscriminatory features. Four filter-based feature ranking algorithms were evaluated, and the SMOTE was applied to balance the class distribution of the dataset in the feature importance calculation. When the Spearman correlation coefficient was used in generating the similarity matrices for hierarchical clustering,

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

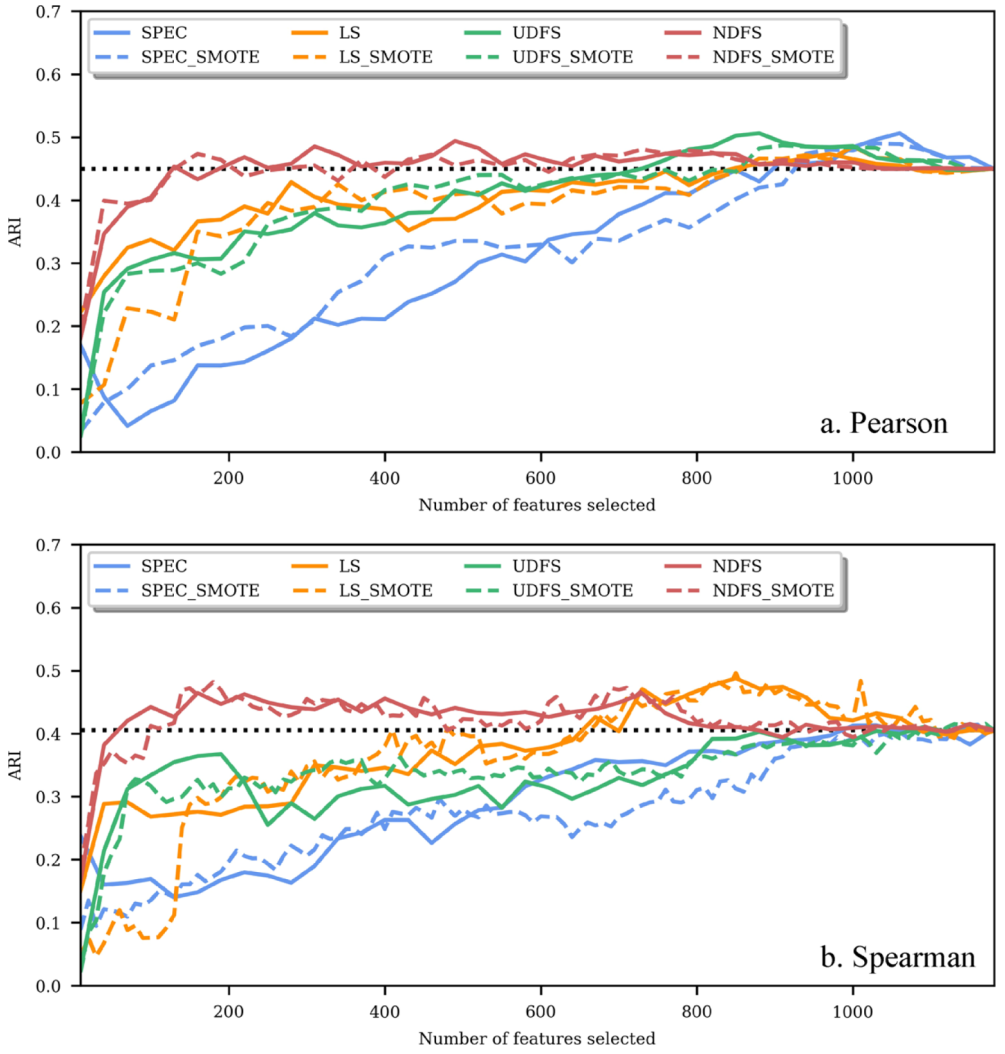


Figure 10. Clustering ARI when increasing the number of top features using different feature selectors. (a) Spectral similarity quantified using the Pearson correlation coefficient. (b) Spectral similarity quantified using the Spearman correlation coefficient.

Table 5. Clustering Results of Different Feature Subsets Using the Spearman Correlation Coefficient for the Affinity Matrix

		ARI (95% CI)			NMI (95% CI)		
	no. features	mean	lower	upper	mean	lower	upper
LS	850	0.470	0.362	0.600	0.623	0.546	0.710
LS_SMOTE	840	0.474	0.365	0.598	0.628	0.550	0.703
NDFS	180	0.460	0.348	0.584	0.623	0.551	0.691
NDFS_SMOTE	180	0.463	0.346	0.582	0.630	0.549	0.708
ensemble1[a]	166	0.465	0.362	0.588	0.626	0.553	0.696
ensemble2[a]	175	**0.472**	0.354	0.601	**0.637**	0.562	0.717
ensemble3[b]	120	0.467	0.361	0.600	0.631	0.555	0.708
ensemble4[c]	228	0.462	0.355	0.584	0.626	0.550	0.701
baseline[d]	1181	0.415	0.309	0.542	0.586	0.504	0.677

[a]Ensembles 1 and 2 include the intersection of the optimal feature subsets from LS and NDFS without or with the SMOTE applied, respectively. [b]Ensemble 3 is the intersection of ensembles 1 and 2. [c]Ensemble 4 is the intersection of the union feature subsets of LS and NDFS. [d]The baseline clustering trees use the full-range features (m = 1181).

the LS and NDFS algorithms were determined to provide the greatest improvement in clustering results. The NDFS feature selector is able to sample the entire spectrum by assigning a high feature importance score to a few features.

The detection method based on IR spectroscopy is based on matching the observed spectrum of an unknown compound to reference spectra of known compounds. Specifically, frequency patterns found in existing spectral databases corresponding to known chemical structures can be matched to measured spectra to identify unknown compounds. The present work shows DFT-calculated IR spectra of a set of known NPS compounds, which complements the knowledge gained from laboratory measurements. Spectra calculated by DFT were shown to retain the most discriminative spectral features when clustering NPS com-

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

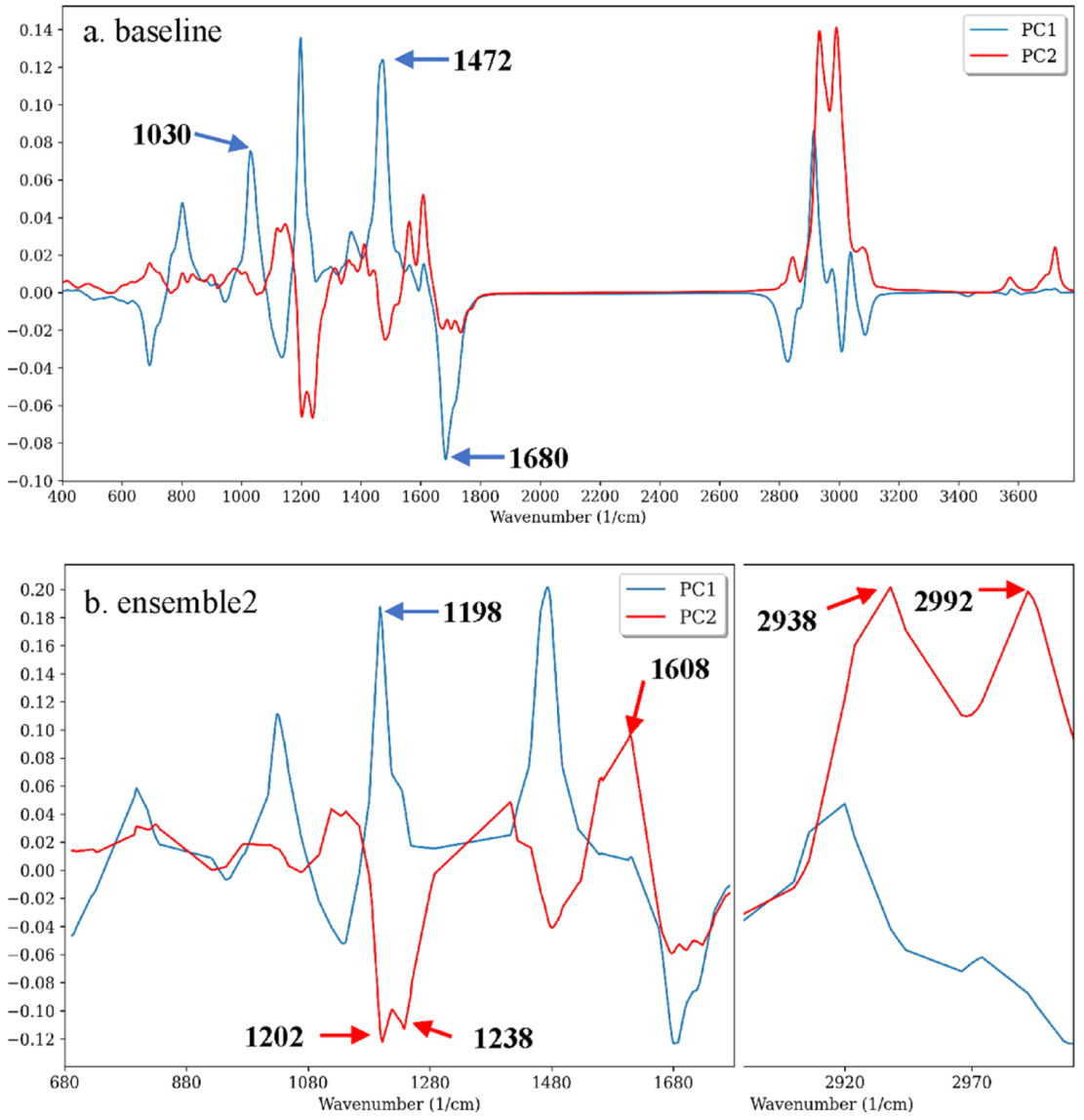


Figure 11. First and second PC loading plot using the full range and feature-reduced datasets. (a) PC1 and PC2 loading plots use the baseline dataset (m = 1181). (b) PC1 and PC2 loading plots use the ensemble 2 subset dataset (m = 175).

pounds based on their pharmacological categorization. This proof-of-concept methodology can be used in filter algorithms to detect spectral features associated with novel designer drugs.

METHODS

Calculation of IR Spectra. According to the UNODC report up to December 2020, the majority of synthetic NPS compounds are stimulants, followed by synthetic cannabinoid receptor agonists and psychedelics with a notable increase in synthetic opioids.[1] This classification is broadly defined according to their pharmacological targets: (1) stimulants mediate the actions of dopamine, norepinephrine, and/or serotonin as reuptake transporter inhibitors;[31−35] (2) cannabinoids primarily interact with G protein-coupled receptors;[36] (3) serotonergic psychedelics are mainly mediated by 5-HT_{2A} receptor agonism;[37,38] and (4) synthetic opioids and fentanyl analogues interact with G protein-coupled opioid receptors as partial to full agonists.[39] A total of 127 unique NPS compounds were selected from 16 major core chemical structure categories. These include 17 natural or synthetic opioids, 62 stimulants (piperidines, tropane alkaloids, amphetamines, cathinones, aminoindanes, and benzofurans), 35 hallucinogens (2C, 2C-B, and 2C-T series and tryptamine), 6 sedatives (benzodiazepines), and 7 cannabinoids. A total of 10 conformers SDF files were downloaded from PubChem for each compound.[40] PubChem3D provides low-energy conformers from a conformer model that samples the energetically accessible and (potentially) biologically relevant conformations of chemical structures using the average atomic pair-wise rmsd.[41] The geometry optimizations were performed using the Gaussian 16 program[42] using the B3LYP level of DFT in combination with the 6-311++G(d, p) basis set. Redundant conformers converged to the same structure were eliminated from the dataset. The harmonic vibrational wavenumbers of all conformers were determined at the corresponding optimized structures, which were confirmed to be local minima by checking that there were no imaginary frequencies. To offset the systematic errors due to basis set incompleteness, neglect of anharmonicity, and incomplete treatment of electron correlation, single scaling factors were applied.[43] To validate the quality of the DFT-simulated IR

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

spectra and the dependence on the basis set, four basis sets [6-31G(d), 6-31+G(d, p), 6-31++G(d, p), and 6-311++G(d, p)] were used, and scaling factors were chosen from the NIST database.[24] The theoretical vibrational frequencies and intensities were convoluted with a Lorentzian distribution, centered at the frequency and multiplied by the intensity. The full width at half-maximum of each distribution was set to 24 cm^{-1} on the basis of the estimated bandwidth observed in the NIST database.[44,45] Finally, the resulting spectra were normalized with respect to the area under the curve and scaled from 0 to 1 and truncated from 400 to 4000 cm^{-1} with a 2 cm^{-1} interval. Quasi-constant features were further removed using *VarianceThreshold* in the *sklearn* package with a threshold value of 9.88×10^{-6}, excluding wavenumbers in 1854−2686 and 3188−3386 cm^{-1} ranges and wavenumbers above 3786 cm^{-1}. The final dataset is with size $n = 930$ and $m = 1181$.

The total IR spectrum is dependent on the temperature results from the contributions of all low-energy conformers, and their contribution was weighted according to their relative population. The total IR spectrum ($\sum$IR) is constructed for each NPS compound using its scaled conformer IR spectrum and Gibbs free energy. The relative populations of the low-energy conformers are computed through the probabilities defined as

$$P(T) = \frac{e^{-\beta \Delta G^k}}{\sum e^{-\beta \Delta G^k}} \quad (1)$$

where $\beta = 1/k_B T$, and k_B is the Boltzmann constant, T is the temperature in Kelvin, and ΔG is the Gibbs free energy of the kth conformer.

Spectrum and Chemical Structure Similarity Measure. The similarity between two spectra, represented by vectors x_A and x_B, is characterized by two statistical measures. Pearson's product moment correlation coefficient based on mean-centered intensities

$$r = \frac{\sum_i (x_{A,i} - \bar{x}_A)(x_{B,i} - \bar{x}_B)}{\sqrt{\sum_i (x_{A,i} - \bar{x}_A)^2} \sqrt{\sum_i (x_{B,i} - \bar{x}_B)^2}} \quad (2)$$

where $x_{A,i}$ and $x_{B,i}$ are the elements of the intensity vectors representing the spectra under comparison and $\bar{x}_A$ and $\bar{x}_B$ are the mean intensity values of spectra A and B, respectively. Spearman's rank correlation coefficient is based on the mean of the intensity differences. A vector d is the difference between the ranks of $x_{A,i}$ and $x_{B,i}$ in their respective dataset

$$\rho = 1 - \frac{6\sum_i di^2}{n \cdot (n^2 - 1)} \quad (3)$$

where n is the number of elements in each vector. Both correlation coefficients range from −1 to +1. A positive Pearson score indicates that all data points are linearly associated, whereas two variables are monotonically related in Spearman correlation even if their relationship is not linear.

The degree to which two molecules are considered "similar" depends on both their structural encoding and the similarity metric used. NPS compounds are often classified based on their pharmacological action and then further classified based on their common chemical scaffold, such as phenethylamines, piperazines, cathinones, tryptamines, and so forth.[46] It is impractical to manually assign labels to NPS compounds as the number of them grows, as does their structural complexity. Molecular fingerprints that encode the molecular structure as binary bit strings allow rapid scan for structural similarity/diversity using a bit-wise comparison on pairs of molecules. The Tanimoto coefficient is a widely used metric for molecular structural similarity quantification.[47] Two types of 2D molecular fingerprints were used in this study. The molecular access system (MACCS) is a structural key fingerprint that encodes for the absence (0) and the presence (1) of a particular structural fragment, with the most commonly used being 166-bit long.[48] The Morgan fingerprint is a circular fingerprint belonging to the extended connectivity fingerprint (ECFP) family that encodes heavy atoms into multiple circular layers up to a given diameter.[49] The RDKit implementation of Morgan with radius = 2 is roughly equivalent to ECFP4. An extended three-dimensional fingerprint (E3FP) is motivated by the ECFP that draws concentrically larger shells and encodes the 3D atom neighborhood patterns from small to larger shells iteratively.[50] Lastly, the MCS similarity is calculated by identifying the structural overlap by matching atomic elements and bond types.[51] Tanimoto values calculated using binary fingerprints will always have a value between 0 and 1, with 1 indicating identical and 0 indicating entirely different. Pertinent details can be found in the Supporting Information. All fingerprints were calculated using *RDKit*.[52]

Clustering Performance Measurement. The full dataset in the ($n = 930$, $m = 1181$) dimension was transformed into (n,n) similarity matrices using Pearson's or Spearman's correlation coefficient, where the value in the ith row and the j-th column indicates the spectral similarity between samples i and j, and each sample is described by its similarity compared to all other samples. The similarity matrices were used as input and submitted to a ward linkage clustering with Euclidean distance as the similarity metric for hierarchical clustering.

The silhouette score is an internal index in measuring the quality of a partition without external information. The optimal number of clusters K was determined by silhouette index analysis,[15] which is a measure of how well cluster members belong to their respective clusters, averaged over all samples

$$SI_k = \frac{1}{n}\sum_{i=1}^{n} \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (4)$$

where n is the total number of points, a_i is the average distance between point i and all other points in its own cluster, and b_i is the minimum of the average dissimilarities between i and points in other clusters.

External indices measure the similarity between the output of the clustering algorithm and the correct partitioning of the dataset. Different clustering trees were compared with each other using the ARI.[16] Let $U = \{u_1, u_2, ..., u_R\}$ and $V = \{v_1, v_2, ..., v_C\}$ represent the external cluster label and that determined by the cluster algorithm, respectively, and n_{ij} is the number of objects belonging to both subsets, u_R and v_j; the ARI is calculated as

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \frac{\left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right]}{\binom{n}{2}}} \quad (5)$$

When two sets of cluster labels have a perfect one-to-one correspondence, the ARI equals unity. The NMI quantifies the

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

mutual dependence between two random variables based on concepts of information theory[17]

$$\mathrm{NMI}(C_i, C_j) = \frac{I(C_i, C_j)}{\sqrt{[H(C_i), H(C_j)]}} \quad (6)$$

where C_i and C_j are cluster assignments of the points generated from feature subsets of feature selectors i and j, respectively. Mutual information $I(C_i,C_j)$ is given as $H(C_i) - H(C_i|C_j)$. $H(C)$ is the Shannon entropy of C, and $H(C_i|C_j)$ is the conditional entropy of C_i, given C_j. NMI = 0 means that two partitions contain no information about one another, whereas NMI = 1 indicates that two partitions contain perfect information about one another.

All hierarchical clusterings are generated using the *fcluster* and *dendrogram* in the *scipy. cluster. hierarchy* package. Heatmaps are generated using the *seaborn* package, The SI, ARI, and NMI are computed using the *sklean. metric* package.

Filter Feature Selection Models and Ensemble Method. Four filter feature selection models were used to rank the features according to certain criteria. The SPEC algorithm studies how to select features according to the structure of the adjacency matrix W and graph G induced from the samples' similarity matrix S.[18] The similarity matrix S is calculated using the radial-base function as a similarity function between two samples, x_i and x_j

$$S_{ij} = \mathrm{e}^{-\left\|x_i - x_j\right\|^2 / 2\sigma^2} \quad (7)$$

The main idea behind SPEC is that the features consistent with the graph structure are assigned similar values to instances that are near to each other in the graph. Therefore, these features should be relevant since they behave similarly in each similar group of samples.[14] The LS is a special case of SPEC that selects the features most consistent with the Gaussian Laplacian matrix using a different ranking function and very efficient with respect to the data size.[19] The UDFS algorithm simultaneously exploits the discriminative information and feature correlation to select discriminative features in the batch mode.[20] Lastly, the NDFS algorithm utilizes spectral clustering to obtain cluster label indicators and a non-negative constraint into the objective function.[21] The sparse feature selection matrix is formulated as an $l_{2,1}$-norm minimization term and solved iteratively. All four feature filter models were implemented in the *scikit-feature* package.[53]

The ensemble method used in this study can be summarized as follows:

1. Rank all the features using each feature selector.
2. Evaluate the clustering performance by increasing the size of top-ranking feature subsets.
3. Identify the feature subsets corresponding to the optimal clustering performance using each feature selector.
4. Identify the common overlap (intersect or union) feature ensembles using different feature selectors.

Model Training, Validation, and Performance Evaluation. In this work, we used a class-imbalanced dataset to reflect the distribution of NPS in the real-world market. Changes in supply, manufacturing, and regulatory regulations all have an impact on the market's continuously moving trend. Machine learning on class-imbalanced data, on the other hand, is biased in favor of the majority class, which is compounded by the high dimensionality of the feature space. The SMOTE is a popular oversampling technique that produces class-balanced data. In this study, we also looked into whether applying the SMOTE improves clustering by computing the feature importance score with and without the SMOTE. The *SMOTE* from *imblean.over_sampling* package was used.

The feature importance ranking was calculated by dividing the dataset by 10-fold and averaging the results over 10 iterations. Each iteration calculates the feature importance score using 90% of the dataset and is repeated 10 times using a different 90% of the dataset. The scores were standardized to a range of 0−1 before being averaged for ease of comparison. The arithmetic mean of the scores acquired from 10 iterations was used to establish the overall feature ranking. Similarly, the clustering evaluations were repeated five times, with a 5-fold split of the dataset for each feature subset chosen using different selectors, and the average ARI and NMI were calculated. The 95% confidence intervals of the ARI and NMI were determined by 250 bootstrap iterations using a sample that is 15% of the size of the dataset for the final ensemble feature subset comparison.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c04945.

DFT IR spectrum simulation using Lorentzian broadening, Boltzmann-weighted total IR spectrum, and hierarchical clustering analysis and comparison (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Kedan He − *Department of Physical Sciences, School of Arts and Sciences, Eastern Connecticut State University, Willimantic, Connecticut 06226, United States;* orcid.org/0000-0002-1953-4956; Phone: 860-465-5759; Email: hek@easternct.edu

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c04945

Funding

Computational resources were provided in part by the MERCURY consortium (http://mercuryconsortium.org/) under NSF grants CHE-1229354, CHE-1662030, and CHE-2018427.

Notes

The author declares no competing financial interest.

■ ACKNOWLEDGMENTS

Clemson University is acknowledged for generous allotment of computation time on the Palmetto cluster.

■ NOMENCLATURE

D, dataset
n, sample size
m, number of features
x_j, jth sample
f_i, ith feature
F, selected feature set
l, number of selected features
K, number of clusters
C_k, kth cluster

REFERENCES

(1) United Nations Office on Drugs and Crime. Early Warning Advisory on New Psychoactive Substances. What are NPS?. https://www.unodc.org/LSS/Home/NPS (accessed Mar 2021).

(2) *"Title 21 United States Code (USC) Controlled Substances Act" United States Drug Enforcement Administration*. https://www.dea.gov/controlled-substances-act. (accessed Mar 2021).

(3) Roggo, Y.; Chalus, P.; Maurer, L.; Lema-Martinez, C.; Edmond, A.; Jent, N. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharm. Biomed. Anal.* **2007**, *44*, 683−700.

(4) Wadood, S. A.; Boli, G.; Xiaowen, Z.; Hussain, I.; Yimin, W. Recent development in the application of analytical techniques for the traceability and authenticity of food of plant origin. *Microchem. J.* **2020**, *152*, 104295.

(5) Zhu, M.-Z.; Wen, B.; Wu, H.; Li, J.; Lin, H.; Li, Q.; Li, Y.; Huang, J.; Liu, Z. The Quality Control of Tea by Near-Infrared Reflectance (NIR) Spectroscopy and Chemometrics. *J. Spectrosc.* **2019**, *2019*, 8129648.

(6) Bel'skaya, L. V. Use of IR Spectroscopy in Cancer Diagnosis. A Review. *J. Appl. Spectrosc.* **2019**, *86*, 187−205.

(7) Luinge, H. J. Automated interpretation of vibrational spectra. *Vib. Spectrosc.* **1990**, *1*, 3−18.

(8) Zloh, M.; Samaras, E. G.; Calvo-Castro, J.; Guirguis, A.; Stair, J. L.; Kirton, S. B. Drowning in diversity? A systematic way of clustering and selecting a representative set of new psychoactive substances. *RSC Adv.* **2017**, *7*, 53181−53191.

(9) Varmuza, K.; Karlovits, M.; Demuth, W. Spectral similarity versus structural similarity: infrared spectroscopy. *Anal. Chim. Acta* **2003**, *490*, 313−324.

(10) Muegge, I.; Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discovery* **2016**, *11*, 137−148.

(11) Yun, Y.-H.; Liang, Y.-Z.; Xie, G.-X.; Li, H.-D.; Cao, D.-S.; Xu, Q.-S. A perspective demonstration on the importance of variable selection in inverse calibration for complex analytical systems. *Analyst (Cambridge, U. K.)* **2013**, *138*, 6412−6421.

(12) Gemperline, P. J.; Salt, A. Principal components regression for routine multicomponent UV determinations: A validation protocol. *J. Chemom.* **1989**, *3*, 343−357.

(13) Geladi, P.; Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(14) Alelyani, S.; Tang, J.; Liu, H. *Feature Selection for Clustering: A Review*, 1st ed.; Chapman and Hall/CRC, 2013; pp 29−60.

(15) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.

(16) Hubert, L.; Arabie, P. Comparing partitions. *J Classif* . **1985**, *2*, 193−218.

(17) Strehl, A.; Ghosh, J. Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions. *J Mach Learn Res.* **2002**, *3*, 583−617.

(18) Zhao, Z.; Liu, H. Spectral feature selection for supervised and unsupervised learning. *Proceedings of the 24th international conference on Machine learning, Corvalis, Oregon, USA*; Association for Computing MachineryCorvalis, Oregon: USA, 2007; pp 1151−1157.

(19) He, X.; Cai, D.; Niyogi, P. Laplacian score for feature selection. *Proceedings of the 18th International Conference on Neural Information Processing Systems*; British ColumbiaMIT PressCanada: VancouverCanadaVancouver, British Columbia, 2005; pp 507−514.

(20) Yang, Y.; Shen, H. T.; Ma, Z.; Huang, Z.; Zhou, X. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two*; AAAI Press: Barcelona, Catalonia, SpainBarcelona, Catalonia, Spain, 2011; pp 1589−1594.

(21) Li, Z.; Yang, Y.; Liu, J.; Zhou, X.; Lu, H. Unsupervised feature selection using nonnegative spectral analysis. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*; AAAI PressCanada: Toronto, Ontario, CanadaToronto, Ontario, 2012; pp 1026−1032.

(22) Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinf.* **2013**, *14*, 106.

(23) Linstrom, P. J.; Mallard, W. G. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; National Institute of Standards and Technology: Gaithersburg MD, 2021.

(24) Johnson, R. D. *NIST Computational Chemistry Comparison and Benchmark Database*; Release, August 2020; Vol. *21*. NIST Standard Reference Database Number 101.

(25) Fraley, C.; Raftery, A. E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Comput. J.* **1998**, *41*, 578−588.

(26) Maulik, U.; Bandyopadhyay, S. Performance evaluation of some clustering algorithms and validity indices. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1650−1654.

(27) Brohée, S.; van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinf.* **2006**, *7*, 488.

(28) de Souto, M. C.; Costa, I. G.; de Araujo, D. S.; Ludermir, T. B.; Schliep, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinf.* **2008**, *9*, 497.

(29) Pirim, H.; Ekşioğlu, B.; Perkins, A. D.; Yüceer, Ç. Clustering of high throughput gene expression data. *Comput. Oper. Res.* **2012**, *39*, 3046−3061.

(30) Kinnunen, T.; Sidoroff, I.; Tuononen, M.; Fränti, P. Comparison of clustering methods: A case study of text-independent speaker modeling. *Pattern Recognit. Lett.* **2011**, *32*, 1604−1617.

(31) Simmler, L. D.; Rickli, A.; Schramm, Y.; Hoener, M. C.; Liechti, M. E. Pharmacological profiles of aminoindanes, piperazines, and pipradrol derivatives. *Biochem. Pharmacol. (Amsterdam, Neth.)* **2014**, *88*, 237−244.

(32) Rickli, A.; Kopf, S.; Hoener, M. C.; Liechti, M. E. Pharmacological profile of novel psychoactive benzofurans. *Br. J. Pharmacol.* **2015**, *172*, 3412−3425.

(33) Luethi, D.; Kaeser, P. J.; Brandt, S. D.; Krähenbühl, S.; Hoener, M. C.; Liechti, M. E. Pharmacological profile of methylphenidate-based designer drugs. *Neuropharmacology* **2018**, *134*, 133−140.

(34) Luethi, D.; Kolaczynska, K. E.; Docci, L.; Krähenbühl, S.; Hoener, M. C.; Liechti, M. E. Pharmacological profile of mephedrone analogs and related new psychoactive substances. *Neuropharmacology* **2018**, *134*, 4−12.

(35) Simmler, L.; Buser, T.; Donzelli, M.; Schramm, Y.; Dieu, L.-H.; Huwyler, J.; Chaboz, S.; Hoener, M.; Liechti, M. Pharmacological characterization of designer cathinones in vitro. *Br. J. Pharmacol.* **2013**, *168*, 458−470.

(36) Mackie, K. Cannabinoid receptors as therapeutic targets. *Annu. Rev. Pharmacol. Toxicol.* **2006**, *46*, 101−122.

(37) Rickli, A.; Moning, O. D.; Hoener, M. C.; Liechti, M. E. Receptor interaction profiles of novel psychoactive tryptamines compared with classic hallucinogens. *Eur. Neuropsychopharmacol.* **2016**, *26*, 1327−1337.

(38) Rickli, A.; Luethi, D.; Reinisch, J.; Buchy, D.; Hoener, M. C.; Liechti, M. E. Receptor interaction profiles of novel N-2-methoxybenzyl (NBOMe) derivatives of 2,5-dimethoxy-substituted phenethylamines (2C drugs). *Neuropharmacology* **2015**, *99*, 546−553.

(39) James, A.; Williams, J. Basic opioid pharmacology—an update. *Br. J. Pain* **2020**, *14*, 115−121.

(40) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388−D1395.

(41) Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem3D: a new resource for scientists. *J. Cheminf.* **2011**, *3*, 32.

(42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe,

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165

D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*: Wallingford, CT, 2016.

(43) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies:An Evaluation of Hartree−Fock ,Møller−Plesset ,Quadratic Configuration Interaction ,Density Functional Theory ,and Semiempirical Scale Factors. *J. Phys. Chem.* **1996**, *100*, 16502−16513.

(44) Mott, A. J.; Rez, P. Calculated infrared spectra of nerve agents and simulants. *Spectrochim. Acta, Part A* **2012**, *91*, 256−260.

(45) Chu, P. M.; Guenther, F. R.; Rhoderick, G. C.; Lafferty, W. J. The NIST Quantitative Infrared Database. *J. Res. Natl. Inst. Stand. Technol.* **1999**, *104*, 59−81.

(46) Luethi, D.; Liechti, M. E. Designer drugs: mechanism of action and adverse effects. *Arch. Toxicol.* **2020**, *94*, 1085−1133.

(47) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, 20.

(48) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Model.* **2002**, *42*, 1273−1280.

(49) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(50) Axen, S. D.; Huang, X.-P.; Cáceres, E. L.; Gendelev, L.; Roth, B. L.; Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. *J. Med. Chem.* **2017**, *60*, 7393−7409.

(51) Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 937−950.

(52) *RDKit: Open-source cheminformatics*. http://www.rdkit.org.

(53) Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. Feature Selection: A Data Perspective. *ACM Comput. Surv.* **2018**, *50*, 1−45 Article 94.

https://doi.org/10.1021/acsomega.1c04945
ACS Omega 2021, 6, 32151−32165